EHUskaratuak: un corpus de traducciones académicas en una lengua minoritaria

[*EHUskaratuak*: a corpus of academic translations in a minority language]

GIDOR BILBAO TELLETXEA Y JESUS MARI MAKAZAGA EIZAGUIRRE
Universidad del País Vasco - Euskal Herriko Unibertsitatea
gidor.bilbao@ehu.es

Resumen

EHUskaratuak (<http://ehuskaratuak.ehu.es>) es un corpus multilingüe que permite consultar textos paralelizados de libros académicos traducidos del inglés, francés y castellano al euskera bajo la coordinación del Servicio de Euskera de la UPV/EHU. Todas las traducciones han sido realizadas aplicando una metodología colaborativa, en la que un profesional de la traducción realiza una primera versión, que es posteriormente consensuada por un grupo de revisores técnicos especialistas del área de conocimiento de la publicación, junto con otro grupo de revisores lingüísticos, especialistas de la lengua meta, el euskera, una lengua minoritaria en proceso de normalización. La aplicación busca ofrecer al usuario los resultados más completos por medio de una búsqueda muy simple, de modo que pueda servir no sólo a profesionales de la traducción o a investigadores, sino también a los propios especialistas universitarios, tanto en la traducción como en la creación de sus propios textos en una lengua en proceso de normalización.

Abstract

EHUskaratuak (<http://ehuskaratuak.ehu.es>) is a multilingual corpus containing Basque translations of academic books in English, French and Spanish which can be consulted alongside their original text. These translations are coordinated by the Basque Normalization and Translation Department of the University of the Basque Country (UPV/EHU). All translations are carried out according to a collaborative methodology: first, a professional translates the text, which is subsequently revised by a group of technical reviewers specialized in the field of knowledge of the publication, and then by a team of copy editors specialized in the target language, Basque, a minority language in the process of normalization. This application seeks to offer the user the most complete results by means of a very simple search, so that it can be useful not only to translation professionals or researchers, but also to university specialists themselves, both in the translation and creation of their own texts in Basque, a language on the way to normalization.

Palabras clave

Corpus multilingüe, textos paralelizados, lenguaje científico-técnico, traducción técnica y normalización de lenguas

Keywords
Multilingual corpus, parallel texts, scientific-technical language, technical translation and language normalization

1. *EHUskaratuak*: el corpus de un servicio de traducción

El nombre con el que bautizamos el corpus de traducciones académicas que presentamos en este artículo encierra en sí un juego de palabras que pretende ser sugerente para el usuario vascófono: *euskara* significa 'lengua vasca'; el verbo derivado *euskaratu* quiere decir 'traducir a la lengua vasca', y finalmente el participio nominalizado *euskaratuak* es utilizado para referirse a 'las cosas traducidas a la lengua vasca', 'las traducciones en lengua vasca'. *Euskaratuak* sería una palabra perfectamente comprensible para cualquier conocedor del vascuence, pero la *-h-* intercalada la convierte en una palabra nueva, inexistente en el léxico de la lengua vasca, que recuerda las siglas en euskera (*EHU*) del nombre oficial bilingüe de la universidad en la que se ha desarrollado el corpus: «Universidad del País Vasco - Euskal Herriko Unibertsitatea (UPV/EHU)». De este modo, si *euskara* es la lengua vasca, *EHUskara* sugeriría la lengua vasca de la UPV/EHU, y *EHUskaratuak* sería, de alguna manera, el conjunto de lo que se ha traducido al euskera en la UPV/EHU[1].

Esta explicación pretende situar el corpus en el contexto en el que fue gestado. De hecho, *EHUskaratuak* (<http://ehuskaratuak.ehu.es>) nació en el seno del Servicio de Euskera de la UPV/EHU[2], un servicio universitario que, entre otras funciones, incluye las de un servicio de traducción. En el caso de la Universidad del País Vasco, este servicio distingue claramente el servicio de traducción de textos administrativos, no académicos, para lo cual se cuenta con una plantilla de traductores pertenecientes al Personal de Administración y Servicios de la universidad, y, por otra parte, un servicio lingüístico que se ocupa fundamentalmente de garantizar la idoneidad y la calidad de la lengua vasca usada en los textos académicos, de los cuales una buena parte son traducciones.

En este punto, quizás no sobre una aclaración en relación con la situación de la lengua vasca en la Universidad del País Vasco. La UPV/EHU reconoce en sus estatutos (artículo 8) la oficialidad del euskera y del castellano, y establece que «prestará particular atención a la cultura vasca y, dentro de ella, al euskera, especialmente en sus aspectos científicos y técnicos». Esta salvaguarda, lejos de reflejar una posición privilegiada de la lengua vasca en la UPV/EHU, supone implícitamente el reconocimiento de su situación no normalizada, y aboga explícitamente por impulsar su normalización en el ámbito universitario. De hecho, en 1990 se aprobó un I Plan de Normalización del Uso del Euskera en la UPV/EHU, que fue seguido en 1999 por el segundo plan con la misma denominación; los planes quinquenales posteriores, aprobados el curso 2007-2008 y el

curso 2013-2014, se llamaron, un tanto eufemísticamente, Plan Director del Euskera de la UPV/EHU.

Unos pocos datos recogidos en el *II Plan Director del Euskera de la UPV/EHU (2013-2017)* nos pueden valer para ilustrar la situación de la lengua vasca como lengua minoritaria en el ámbito universitario, aún tomando en consideración que, teniendo en cuenta tanto universidades públicas como privadas, la UPV/EHU es la universidad que indudablemente ha realizado un mayor esfuerzo por impulsar el desarrollo del euskera tanto en la docencia universitaria como en la investigación.

Si examinamos los datos del alumnado, nos parece significativo observar que en el curso 1997/98 cursaba sus estudios en los grupos de euskera el 23,46 % del alumnado de nuevo acceso, y en el curso 2012/13 el 48,07 %. Si lo comparamos con los porcentajes de profesorado, observamos que en el año 2013 el 47,76 % del PDI de plantilla con dedicación completa ocupaba una plaza de perfil lingüístico bilingüe (español-euskera), de lo cual se puede extrapolar que cerca de la mitad del profesorado no sabe euskera (el dato no es exacto, porque hay profesores que saben euskera y ocupan puestos no bilingües, pero no son numerosos). Finalmente, es interesante el dato relativo a la docencia en lengua vasca, y su incremento en un periodo no muy amplio: teniendo en cuenta que, en general, no hay titulaciones ni asignaturas que se oferten exclusivamente en lengua vasca, se ha pasado de ofertar en 1988 el 32 % de las asignaturas obligatorias en euskera, a un 80 % en 2013. Esta evolución ha ido acompañada por un esfuerzo en la creación y traducción de material bibliográfico en lengua vasca, pero el déficit es aún hoy evidente.

Es precisamente en este aspecto donde interviene el Servicio de Euskera de la UPV/EHU, que cuenta entre sus funciones la coordinación y el control de la calidad de tres colecciones de traducciones académicas: 1) Una colección de manuales universitarios, que edita el Servicio Editorial de la Universidad del País Vasco. 2) Una colección de prosa de divulgación científica (*Zientzia Irakurle Ororentzat* 'Ciencia para todos los lectores'), que edita junto con la Cátedra de Cultura Científica de la UPV/EHU. 3) Una colección bilingüe de leyes (llamada *Legeak - Leyes*), que la universidad edita en colaboración con el Departamento de Justicia del Gobierno Vasco.

Al menos desde 2002, el Servicio de Euskera de la UPV/EHU ha enmarcado todos sus proyectos de traducción en un proyecto de normalización de la lengua vasca académica, gracias a la clarividencia de un precursor en el campo de los corpus de la lengua vasca y en general de las herramientas tecnológicas aplicadas al estudio del euskera como fue el profesor Iñaki Ugarteburu.

Este proyecto, en su primer diseño, de hace ya más de una docena de años, incluía una herramienta de consulta online de las memorias de traducción de todos

los textos académicos traducidos bajo la coordinación y supervisión del Servicio; aquella herramienta (*Itzulpenen kontsulta* 'Consulta de traducciones') fue de hecho, en 2002, el antecedente directo, organizado según la tecnología disponible en aquel momento, de lo que luego ha sido, a partir de 2013, *EHUskaratuak*.

En este contexto hay que mencionar también *Ereduzko Prosa Gaur*, el corpus de la lengua vasca más exitoso hasta el momento, como veremos más adelante, que fue publicado por el Servicio de Euskera de la UPV/EHU en su página web en 2007, fundamentalmente gracias al trabajo del profesor Ibon Sarasola, miembro del Servicio.

Volviendo al proyecto ideado por Iñaki Ugarteburu, la consulta del texto de las traducciones se completaba con una base de datos terminológica que, desgraciadamente, pronto dejó de actualizarse, con lo cual perdió muy pronto gran parte de su valor. Ahora, en 2014, estamos a punto de publicar la base de datos terminológica *Gaika* ('Por temas', 'Por áreas'), relacionada por medio de *links* con el corpus *EHUskaratuak*, puesto que recoge precisamente la terminología utilizada en dicho corpus.

Finalmente, el tercer componente del proyecto de normalización lo constituía y sigue constituyendo *Ehulku* ('Consejos lingüísticos de la UPV/EHU'), un servicio de consulta de dudas lingüísticas atendido por miembros del Servicio de Euskera, completado por una serie de cápsulas de formación lingüística que el profesorado vascófono de la UPV/EHU recibe regularmente en su dirección de correo electrónico.

Insistimos en que estas tres herramientas fueron concebidas para optimizar el valor de las traducciones académicas coordinadas y supervisadas por el Servicio de Euskera. El denominador común de las tres colecciones de traducciones mencionadas más arriba es la metodología colaborativa en la traducción, ya que en todas las traducciones colaboran desde el inicio mismo de la traducción uno o varios profesionales de la traducción, uno o varios revisores técnicos especialistas del área de conocimiento del libro traducido, y un grupo de revisores lingüísticos especialistas de la lengua meta (el euskera). Todos ellos estudian, antes de comenzar a elaborar una primera versión de la traducción, cuáles son los problemas terminológicos, fraseológicos... específicos del texto que debe ser traducido, y cada uno desde su conocimiento contribuye a que la traducción sea fiel al original, natural desde el punto de vista de la lengua meta, y correcta desde la perspectiva del especialista de la materia.

No podemos ocultar que todos estos proyectos suponen a la UPV/EHU un esfuerzo económico cuantioso. Sólo en la partida presupuestaria de apoyo a la producción y traducción de libros universitarios de texto, es decir, fundamentalmente en la primera de las colecciones citadas, la universidad viene invirtien-

do estos últimos años (manejamos datos de 2011, 2012 y 2013) en torno a 500.000 euros anuales, sufragados a través de contratos programa financiados por el Gobierno Vasco.

Una inversión de tal envergadura exige a los responsables del proyecto un esfuerzo suplementario para hacer visibles y útiles los productos obtenidos, y es precisamente en este contexto donde los autores de este artículo, desde los puestos de responsabilidad que ocupaban en cada momento, y partiendo desde su propia experiencia como traductores e investigadores en proyectos como *EGLU* (la gramática de la Real Euskaltzaindia de la Lengua Vasca, en la que Makazaga es coautor desde los primeros volúmenes), *ZTHB* (el diccionario de términos científicos y técnicos de la Real Euskaltzaindia de la Lengua Vasca, en el que participa Bilbao), el servicio de consulta de dudas lingüísticas *Ehulku* en el que colabora Makazaga, o incluso el proyecto de edición de textos *Monumenta Linguae Vasconum* del cual es partícipe Bilbao, cobran conciencia de la importancia creciente de los corpus como herramientas de consulta imprescindible en la toma de decisiones de carácter lingüístico, y en 2012 llegan a la conclusión de que es el momento de recoger y publicar en un corpus online paralelo todas las traducciones coordinadas y supervisadas por el Servicio de Euskera de la UPV/EHU.

2. Los corpus públicos de la lengua vasca

Antes de presentar nuestro corpus, merece la pena presentar una visión panorámica de los corpus públicos de la lengua vasca a día de hoy. Lo hacemos pensando que puede ofrecer una aproximación válida al lector que se acerca a este artículo desde un conocimiento lejano de la situación de la lengua vasca y de los trabajos desarrollados en torno a la lengua vasca; y por otra parte, desde la convicción de que, dada la rapidez con la que evoluciona el estado de la cuestión en relación con este tipo de herramientas, es interesante contar periódicamente con fotos fijas que dejen constancia de la realidad concreta en un momento determinado, en la línea de un trabajo presentado por Miriam Urkia con ocasión de la publicación del corpus *Consumer corpusa* (Urkia 2010), o incluso de uno anterior, de 2008, obra de varios colaboradores de la Fundación Elhuyar (Areta, Gurrutxaga y Leturia 2008).

Recogemos en la «Tabla 1» los corpus públicos de la lengua vasca disponibles en la fecha actual (mayo de 2014). En cada caso, recogemos en cursiva el nombre del corpus, junto a cada uno de los cuales añadimos entre comillas simples nuestra traducción (que en muchos casos puede servir para ofrecer una idea aproximada del contenido) y entre paréntesis angulares la dirección en la que está disponible. En cada caso definimos luego cada corpus según los parámetros que nos han parecido interesantes para hacer después una presentación sintética

del panorama: los focos de producción de los corpus de la lengua vasca, si se trata de un corpus histórico o actual, si es monolingüe, bilingüe o multilingüe, si los textos recogidos corresponden a lenguajes de especialidad o a la lengua general, si se trata de lengua traducida o lengua no traducida, y, finalmente, si el corpus en cuestión está lematizado o no.

OEH-ko Testu Corpusa 'Corpus de textos del Diccionario General Vasco' (CD)					
Euskaltzaindia	Histórico	Monolingüe	Gen.-Liter.	No traducido	No lemat.
XX. mendeko euskararen corpus estatistikoa 'Corpus estadístico de la lengua vasca del siglo XX' <http://xxmendea.euskaltzaindia.net/Corpus/>					
Euskaltzaindia	Actual	Monolingüe	General	No traducido	Lematizado
Lexikoaren Behatokiaren corpusa 'Corpus del Observatorio del Léxico' <http://lexikoarenbehatokia.euskaltzaindia.net/cgi-bin/kontsulta.py>					
Euskaltzaindia	Actual	Monolingüe	General	No traducido	Lematizado
EHUskaratuak 'Traducciones de la UPV/EHU' <http://ehuskaratuak.ehu.es/kontsulta/>					
UPV/EHU	Actual	Bilingüe	Científ.-técn.	No traducido	Lematizado
Egungo Testuen Corpusa (ETC) 'Corpus de Textos Actuales' <http://www.ehu.es/etc/>					
UPV/EHU EI	Actual	Monolingüe	General	No traducido	(No) lemat.
Ereduzko Prosa Gaur (EPG) 'Prosa de referencia actual' <http://www.ehu.es/euskara-orria/euskara/ereduzkoa/>					
UPV/EHU EI	Actual	Monolingüe	Literario	No tr. + Tr.	(No) lemat.
Ereduzko Prosa Dinamikoa (EPD) 'Prosa de referencia dinámica' <http://ehu.es/ehg/epd/>					
UPV/EHU EI	Actual	Monolingüe	General	No traducido	(No) lemat.
Euskal Klasikoen Corpusa (EKC) 'Corpus de clásicos vascongados' <http://www.ehu.es/ehg/kc/>					
UPV/EHU EI	Histórico	Monolingüe	Literario	No traducido	(No) lemat.
Goenkale Corpusa 'Corpus de la serie televisiva *Goenkale*' <http://ehu.es/ehg/goenkale/>					
UPV/EHU EI	Actual	Monolingüe	General	No traducido	(No) lemat.
Pentsamenduaren Klasikoak corpusa 'Corpus de la colección Clásicos del Pensamiento' <http://ehu.es/ehg/pkc/>					
UPV/EHU EI	Actual	Monolingüe	Filosófico	No traducido	(No) lemat.
ZIO corpusa 'Corpus de la colección ZIO de prosa de divulgación científica' <http://www.ehu.es/ehg/zio/>					

UPV/EHU EI	Actual	Monolingüe	Científ.-Lit.	Traducido	(No) lemat.
Zuzenbide Corpusa 'Corpus de derecho' <http://www.ehu.es/ehg/zuzenbidea/>					
UPV/EHU EI	Actual	Monolingüe	Jur.-adm.	Traducido	(No) lemat.
Itzulpen juridikoen kontsulta 'Consulta de traducciones de textos jurídicos' <http://www.justizia.net/euskara-justizian>					
EJ-GV Just.	Actual	Bilingüe	Jur.-adm.	Traducido	No lemat.
Epaitegietako agirien corpus elebiduna 'Corpus bilingüe de documentos de los juzgados' <http://www.justizia.net/uzei/default.asp?idioma=eu>					
EJ-GV Just.	Actual	Bilingüe	Jur.-adm.	Traducido	No lemat.
Bizimena: Bizkaiko Foru Aldundiko itzulpen memoriak 'Modo de vida: memorias de traducción de la Diputación Foral de Bizkaia' <http://www.bizkaia.net/home2/Temas/DetalleTema.asp?Tem_Codigo=6130>					
BFA	Actual	Bilingüe	Jur.-adm.	Traducido	No lemat.
Gipuzkoako Foru Aldundiaren itzulpenen datu basea 'Base de datos de traducciones de la Diputación Foral de Gipuzkoa' <http://www.gipuzkoa.net/imemoriak/>					
GFA	Actual	Bilingüe	Jur.-adm.	Traducido	No lemat.
CorpEus 'CorpEusk, corpus del euskera' <http://corpeus.elhuyar.org/cgi-bin/kontsulta.py>					
Elhuyar	Actual	Monolingüe	General	No traducido	Lematizado
Elhuyarren web-corpus elebakarra 'Corpus-web monolingüe de la Fundación Elhuyar' <http://webcorpusak.elhuyar.org/cgi-bin/kontsulta.py?mota=arrunta>					
Elhuyar	Actual	Monolingüe	General	No traducido	Lematizado
Elhuyarren web-corpus paraleloa 'Corpus-web paralelo de la Fundación Elhuyar' <http://webcorpusak.elhuyar.org/cgi-bin/kontsulta2.py?mota=arrunta>					
Elhuyar	Actual	Bilingüe	General	Traducido	Lematizado
Zientzia eta teknologiaren corpusa 'Corpus de ciencia y tecnología' <http://www.ztcorpusa.net/cgi-bin/kontsulta.py>					
Elhuyar	Actual	Monolingüe	Científ.-técn.	No traducido	Lematizado
Consumer corpusa 'Corpus de la revista *Consumer*' <http://corpus.consumer.es/corpus/>					
Elh. + Eroski	Actual	Multilingüe	General	Traducido	Lematizado

Tabla 1: Corpus públicos de la lengua vasca en 2014

Si nos fijamos en los focos de producción, es evidente que ha sido la universidad pública UPV/EHU la institución que ha realizado hasta el momento el mayor esfuerzo en la creación de corpus de la lengua vasca. Hay que destacar la

labor que en este campo está realizando el instituto de investigación *Euskara Institutua* 'Instituto de Euskera' («EI» en la tabla), en cuya página web se pueden consultar actualmente 8 corpus diferentes; también *EHUskaratuak* está disponible en la página web de la misma universidad, pero, en este caso, no se trata de un proyecto del Instituto de Euskera, sino del Servicio de Euskera.

A la UPV/EHU le sigue en número de corpus producidos y publicados la Fundación Elhuyar, una fundación que dio sus primeros pasos como asociación cultural y lleva más de cuarenta años trabajando para consolidar el euskera en el ámbito de la ciencia y de la tecnología, y que colabora con la UPV/EHU en diversos proyectos; en la página web de Elhuyar podemos consultar cuatro corpus diferentes, a los que hay que añadir un quinto, producido junto con una empresa privada.

También merece una mención especial *Euskaltzaindia*, la Real Academia de la Lengua Vasca, en cuya página encontramos actualmente dos corpus (aunque en realidad uno es continuación del otro) y fue pionera con la publicación de un corpus en formato CD. Finalmente, también otras instituciones públicas como son las diputaciones forales (con el impulso de sus servicios de traducción) o el departamento de Justicia del Gobierno Vasco han hecho sus aportaciones en este campo.

Con la etiqueta de *corpus histórico* hemos querido reunir los corpus de la lengua vasca que reúnen textos anteriores al año 1900. Observamos que tan solo dos de los 21 corpus consultados cumplen ese requisito, y 19 de ellos corresponden, por tanto, a la lengua vasca actual.

Doce de los corpus examinados recogen lenguajes de especialidad: cinco de ellos recopilan textos de lenguaje jurídico-administrativo, tres son de lenguaje científico-técnico, otros tres son textos literarios y uno es exclusivamente filosófico. Hemos etiquetado el resto como corpus de lengua general, aunque en ocasiones resulta difícil distinguir estos de los literarios (es el caso, por ejemplo, de un corpus que recoge los textos escritos de los guiones de una serie de televisión).

Nos ha interesado así mismo saber cuántos corpus recogen textos de lengua traducida, y cuántos corresponden a textos creados en euskera, no traducidos. Sin atender al número de caracteres que recoge cada corpus, podemos decir que son justo mitad y mitad: diez corpus de lengua traducida y diez de lengua no traducida, más uno que reúne textos de los dos tipos, pero dispone de un filtro de búsqueda que permite discernir los unos de los otros.

La mayor parte de los corpus públicos de la lengua vasca son monolingües (catorce), pero la cantidad de corpus bilingües es considerable: cinco de ellos son bilingües y recogen textos paralelos euskera-español, el par de lenguas más

frecuente en la traducción en lengua vasca, y uno (precisamente *EHUskaratuak*) es bilingüe con el euskera formando pareja con el español, el francés o el inglés, dependiendo del texto. Tan solo un corpus es realmente multilingüe y recoge textos paralelos de las cuatro lenguas oficiales del estado español (español, euskera, catalán y gallego).

Finalmente, un parámetro en nuestra opinión fundamental para evaluar la calidad de los corpus de una lengua como el euskera es la lematización. Hay que tener en cuenta que la búsqueda de una secuencia de caracteres como *jario*, en euskera, incluso con el uso de truncadores, podría permitir obtener resultados referidos exclusivamente a la categoría gramatical de «nombre»: *jario, jarioa, jarioak, jarioen, jariorik...* y quedarían excluidos *darizkion, dario, zerien, darion, dariela, zerizkiola, zerien...* que solo podríamos localizar en un corpus lematizado que permitiera distinguir en su búsqueda «*jario* [verbo]» y «*jario* [nombre]». Entre los corpus analizados, cinco de ellos realizan exclusivamente búsquedas de secuencias de caracteres y no disponen de lematizador; ocho se definen como corpus lematizados, y disponen de un lematizador más o menos desarrollado. En el caso de los ocho corpus del Instituto de Euskera de la UPV/EHU, no se puede decir que sean corpus no lematizados, puesto que las búsquedas devuelven resultados que no se esperarían a partir de la secuencia de caracteres; sin embargo, el nivel de lematización es muy básico.

3. *EHUskaratuak*: un corpus de traducciones académicas

Ya hemos adelantado que la idea del proyecto *EHUskaratuak* partió del Servicio de Euskera de la UPV/EHU, como una forma de hacer más visibles y útiles para un público amplio (traductores, terminólogos, lingüistas... por supuesto, pero también profesorado y alumnado universitario de todas las áreas de conocimiento) las traducciones académicas coordinadas y supervisadas por el servicio a lo largo de los años. El desarrollo tecnológico, por su parte, ha sido obra de *Elhuyar. Hizkuntza eta teknologia* 'Elhuyar. Lengua y tecnología', una sección de la Fundación Elhuyar, dedicada específicamente al desarrollo de recursos y herramientas como resultado de I+D, dirigidos a diversas industrias de la lengua, en estrecha colaboración con diferentes departamentos y proyectos de investigación de la UPV/EHU. Cabe señalar que el desarrollo del etiquetado y el sistema de consulta de *EHUskaratuak* permitió la realización de un trabajo de fin de máster en el departamento de Lenguajes y Sistemas Informáticos de la UPV/EHU, en el Máster de Análisis y Procesamiento de la Lengua (Jauregi 2012). Para decidir la asignación del encargo (financiado con dinero de la partida del Vicerrectorado de Euskera y Plurilingüismo dedicada a «compra, localización y mantenimiento de aplicaciones en euskera») fueron criterios decisivos precisamente la colaboración con nuestra universidad en distintos proyectos y la

garantía que ofrecía su lematizador, ya probado en otros recursos desarrollados por la misma empresa.

Centrándonos en los textos que conforman el corpus *EHUskaratuak*, desde el principio pensamos en una fórmula que permitiera un incremento progresivo de la cantidad de textos, hasta llegar a integrar todas las traducciones académicas coordinadas por el Servicio de Euskera. En una primera fase se compilaron los libros publicados entre 2010 y 2012, para ir añadiendo en fases sucesivas, correspondientes a años naturales, la producción de ese año y la del año inmediatamente anterior al más antiguo ya compilado, de modo que la secuencia de las fases sería la siguiente: 2010-2012 > 2009-2013 > 2008-2014 > 2007-2015... En estos momentos el corpus recoge 51 libros publicados entre 2009 y 2013 (más algunos más antiguos, exclusivamente de la colección *ZIO*, de prosa de divulgación científica), de 18 subáreas de conocimiento, que se distribuyen de la siguiente forma (señalamos junto a cada subárea el número de libros recogido en el corpus para cada par de lenguas):

Ciencias de la Vida
Biología: 4 EN > EU
Medicina: 3 EN > EU
Humanidades
Arte: 2 ES > EU; 1 EN > EU
Filología: 1 EN > EU
Filosofía: 1 EN > EU
Historia: 1 ES > EU; 1 EN > EU
Literatura: 1 ES > EU
Ciencias Sociales
Economía: 1 ES > EU
Pedagogía: 7 ES > EU; 1 FR > EU
Psicología: 1 EN > EU
Sociología: 2 EN > EU; 1 ES > EU; 1 FR > EU
Derecho: 1 ES > EU; 1 EN > EU
Ciencias Exactas y de la Materia
Física y Química: 2 ES > EU
Geología: 1 ES > EU; 1 EN > EU
Matemáticas: 1 EN > EU
Ciencia y Tecnología
Arquitectura: 1 ES > EU; 1 EN > EU
Informática: 1 ES > EU; 2 EN > EU
Ingeniería: 7 ES > EU; 3 EN > EU

Tabla 2: Número de libros del corpus *EHUskaratuak* para cada subárea y par de lenguas

Como vemos en la «Tabla 2», el par de lenguas más frecuente es el español-euskera, con 26 libros, seguido de cerca por el par inglés-español, con 23 publicaciones; en el par francés-euskera hay tan solo 2 libros. Sin embargo, si observamos el dato correspondiente al número de frases traducidas en cada lengua, el

resultado varía considerablemente, debido a que, en general, los libros de texto universitarios que el servicio ha traducido del inglés al euskera son mucho más voluminosos que los que ha traducido del español al euskera. La «Tabla 3» es clara en este sentido:

Idioma	Número de frases	Número de palabras
Inglés (lengua origen)	257.947	3.851.363
Español (lengua origen)	155.924	2.915.413
Francés (lengua origen)	9.470	183.727
Euskera (lengua meta)	423.341	5.455.708
TOTAL	846.682	12.406.211

Tabla 3: Número de frases y palabras del corpus *EHUskaratuak*, por lenguas

En este punto, puede resultar interesante explicitar que, para facilitar la compilación de los textos del corpus, en todos los pliegos de condiciones de los encargos de traducción del Servicio de Euskera de los últimos años (a partir de 2008, aproximadamente), se incluye la obligatoriedad de entregar, junto con la traducción, la memoria de traducción correspondiente, pensando precisamente en facilitar la creación de productos derivados como el corpus que nos ocupa. En el caso de traducciones más antiguas, o de algunas traducciones de las que no disponemos de memoria de traducción, nos vemos obligados a realizar el trabajo *a posteriori*, obteniendo el texto alineado a partir de diferentes formatos (en ocasiones disponemos del texto traducido en formato *doc*, pero debemos escanear ahora el original; y, en el caso de las traducciones más antiguas, incluso debemos escanear la propia traducción).

Para comenzar a describir la propia aplicación *EHUskaratuak*, debemos recordar el interés por llegar a un público amplio, no necesariamente especializado en lingüística, lo cual exige desarrollar una interfaz de búsqueda completa y compleja para satisfacer las necesidades del público interesado, pero al mismo tiempo intuitiva y sencilla de manejar, pensada para poder obtener resultados útiles también a través de la búsqueda simple.

El usuario puede elegir entre consultar la interfaz de búsqueda y todos los demás elementos de la aplicación (presentación, ayuda e información sobre textos compilados) en euskera o en español.

Aunque en algunos sitios se ha definido el corpus como corpus multilingüe (Jauregi 2012), porque incluye cuatro lenguas diferentes, nosotros preferimos definirlo como corpus paralelo bilingüe, ya que los textos son paralelos siempre de dos en dos, siendo el euskera siempre lengua meta, mientras que, dependiendo del libro, el inglés, el español o el francés son la lengua origen; si bien es cierto que, si se realiza la búsqueda desde un término en euskera, el usuario puede elegir las diferentes lenguas en las que quiere ver el resultado, o las tres (inglés, español y francés); no obstante, el resultado será siempre una serie de

textos paralelos en diferentes combinaciones de lenguas (inglés > euskera; español > euskera; francés > euskera), siempre de dos en dos.

Por ejemplo, si buscamos en los dos diccionarios bilingües de lengua general más al uso la palabra castellana *fluido*, uno de ellos (*Elhuyar hiztegia*, de la Fundación Elhuyar) propone en euskera *fluido* con la marca de especialidad «Fís.», y la colocación «*mecánica de fluido* : *fluidoen mekanika*». El otro diccionario más usado (*Zehazki*, de Ibon Sarasola), para la misma búsqueda, propone *jariakin* con la citada marca de especialidad «Fis.», y precisamente la misma colocación, pero con una propuesta diferente: «*mecánica de fluido*: *jariakinen mekanika*». La diferencia entre las dos opciones estriba en adoptar para el término de especialidad un préstamo «internacional» de origen latino, o usar un neologismo derivado de la palabra patrimonial *jario*. Si queremos comprobar en *EHUskaratuak* cuál de los dos términos vascos ha sido más utilizado en las traducciones académicas de la UPV/EHU, elegiremos como idioma «Euskera», pediremos buscar «lema» (para que la búsqueda incluya palabras declinadas como *jarioa, jarioen, jarioarekin*), señalaremos la categoría «nom», y nos interesará conocer los resultados en las tres combinaciones de lengua posibles. El resultado es clarificador: *fluido* aparece en 20 libros, con 1413 coincidencias, y *jariakin* en 11 libros, con 48 coincidencias. Además, si nos fijamos en los textos originales, vemos que, en varias ocasiones, el término original es el español *secreción* o el inglés *secretion*.

Como el corpus está etiquetado y lematizado tanto en euskera como en inglés, español y francés, la propia herramienta *EHUskaratuak* nos permitirá proseguir la investigación, rastreando a través de la búsqueda simple las traducciones que se han propuesto para el lema español *fluido* o *secreción*, para el lema inglés *fluid* o *secretion* y para el lema francés *secrétion*. Esta búsqueda nos permitirá comprobar que en el área de ciencias de la vida también se utiliza en euskera *sekrezio* para traducir el inglés *secretion*, o incluso *jario* para el francés *sécrétion*.

La búsqueda avanzada, a su vez, permite realizar la búsqueda en una combinación de lenguas. Siguiendo con el ejemplo anterior, podemos buscar los textos en los que el lema inglés *fluid* ha sido traducido por el lema en euskera *jariakin*. Un estudio detenido de los resultados obtenidos nos permitirá comprobar que todos ellos se refieren al ámbito de las ciencias de la vida (aun cuando el libro haya sido catalogado en alguna otra de las áreas).

El siguiente paso sería comprobar si la colocación *fluidoen mekanika* o la colocación *jariakinen mekanika* han sido usadas en alguna ocasión en las traducciones académicas de la UPV/EHU. Para ello podemos usar también la búsqueda avanzada, eligiendo el idioma «Euskera» en las dos ventanas de búsqueda, y escribiendo el lema *fluido* (y luego *jariakin*) en una de ellas, y el lema *mekanika*

en la otra; debemos especificar la distancia *0* entre ellas. El resultado es también clarificador: nunca se ha utilizado *jariakinen mekanika*, y aparece en cuatro ocasiones *fluidoen mekanika*.

La posibilidad de establecer la distancia entre las dos palabras que buscamos, permite realizar investigaciones más complejas. Por ejemplo, para ver cómo se traducen o formulan las definiciones en euskera, en un contexto académico, podemos realizar las siguientes búsquedas avanzadas, con resultados francamente interesantes:

ESP Lema: *definir*

Categoría: verb.

ESP Forma: *como*

Distancia: 2

ENG Lema: *define*

Categoría: verb.

ENG Forma: *as*

Distancia: 2

En una lengua de las características del euskera, es importante el uso adecuado de los truncadores para buscar palabras compuestas (por ejemplo «**-landaredi*» y «*landaredi-**»), sufijos (por ejemplo «**tasun*») o desinencias (por ejemplo «**kiko*»).

Finalmente, debemos señalar que la aplicación permite generar automáticamente, a partir de los resultados, gráficos estadísticos circulares.

4. *EHUskaratuak*: un corpus para traductores, académicos e investigadores

Más arriba hemos apuntado que *EHUskaratuak* pretende ser una herramienta útil para un público amplio que pueda acercarse a este corpus por diferentes intereses.

Por un lado, pensamos en la utilidad del corpus para traductores que deban traducir textos académicos de características similares a los textos del corpus, y en los académicos, tanto profesores como estudiantes, que deben producir textos académicos en lengua vasca, en un contexto en el que los materiales de referencia no son abundantes, por las circunstancias que hemos tratado de resumir en el primer apartado de este trabajo. En nuestra opinión, para que este tipo de públi-

co acceda regularmente a un corpus, este debe cumplir tres condiciones: 1) El manejo debe ser sencillo e intuitivo. 2) El corpus debe tener un tamaño suficiente, para que la escasez de resultados no produzca frustración. 3) El corpus debe ser conocido y accesible, para lo cual es importante que aparezca en directorios especializados y que establezca alianzas con otros corpus de la misma lengua, para facilitar mutuamente su visibilidad.

Por otro lado, queremos también ofrecer una herramienta útil a los investigadores, tanto de la traducción como de la lengua. En este caso, las exigencias son sensiblemente diferentes. En lugar del manejo intuitivo, el investigador valorará la existencia de filtros de búsqueda adecuados. La visibilidad del corpus tampoco será importante para ellos. Pero también este tipo de usuario, como el anterior, necesitará un corpus de tamaño suficiente para que sus resultados sean representativos.

Pensamos que la herramienta *EHUskaratuak* cumple buena parte de los requisitos que le exige y exigirá la comunidad académica y científica vascófona. El reto actual es seguir creciendo progresivamente, de acuerdo con la planificación inicial, para conseguir que la totalidad de las traducciones académicas coordinadas y supervisadas por el Servicio de Euskera de la UPV/EHU esté efectivamente en el corpus.

5. Bibliografía

Areta, Nerea, Antton Gurrutxaga e Igor Leuria. 2008. Begiratu bat corpus-baliabideei. *@ BAT Soziolinguistika aldizkaria* 66: 71-92. Disponible en <http://www.soziolinguistika.org/files/Nerea %20Areta %2C %20Antton %20 Gurrutxaga %2C %20Igor %20Leturia.pdf >. Consultado el 27/5/2014.

Jauregi, Amaia. 2012. *EHUskaratuak. Corpus eleaniztun baten etiketatze automatikoa eta kontsulta-sistemaren garapena* [Tesis de máster]. Disponible en <http://ixa.si.ehu.es/master/master_tesiak>. Consultado el 27/5/2014.

Urkia, Miriam. 2010. Corpusgintzaren garrantzia hizkuntzalaritzan eta euskararen egoera. *@ Plazaberri* Enero 2010. Disponible en <http://www.euskaltzaindia.es/dok/plazaberri /2010/urtarrila/corpusgintza_miriamurkia.pdf>. Consultado el 27/5/2014.

Vicerrectorado de Euskera de la UPV/EHU. 2014. *II Plan Director del Euskera de la UPV/EHU (2013-2017)*. Bilbao: Servicio Editorial de la Universidad del País Vasco.

[1] Gidor Bilbao ha participado en este artículo en el contexto de un proyecto de investigación financiado por el Ministerio de Ciencia e Innovación de España, dirigido por el profesor Joseba A. Lakarra: «Monumenta linguae Vasconum (IV)» (FFI2012-37696). Este proyecto forma parte

del grupo de investigación consolidado GIC 07/89-IT-473-07, financiado por el Gobierno Vasco, y se incluye en la Unidad de Formación e Investigación UFI11/14 de la UPV/EHU. Jesus Mari Makazaga participa en el proyecto de investigación GIU13/23 de la UPV/EHU.

[2] Uno de los autores, Jesus Mari Makazaga, ha trabajado en este Servicio de Euskera de la UPV/EHU desde 2002, y es su director desde 2011. Gidor Bilbao ha sido director del servicio de octubre de 2007 a junio de 2009 y Vicerrector de Euskera y Plurilingüismo, el vicerrectorado del que depende el servicio, de enero de 2011 a enero de 2013.